

Achievable speedup in C++ CUDA with hand-tuned GPU data transfer and execution overlap

